Federating distributed and heterogeneous information sources in neuroimaging: the NeuroBase Project.

Christian Barillot, Habib Benali, Michel Dojat, Alban Gaignard, Bernard Gibaud, Serge Kinkingnéhun, Jean-Pierre Matsumoto, Mélanie Pélégrini-Issac, Eric Simon, Lynda Temal

Federating Distributed and Heterogeneous Information Sources in Neuroimaging: The NeuroBase Project

C. Barillot[1], H. Benali[2], M. Dojat[4], A. Gaignard[1], B. Gibaud[1], S. Kinkingnéhun[2], J-P. Matsumoto[5], M. Pélégrini-Issac[2], E. Simon[5], L.Temal[1]

Federating Distributed and Heterogeneous Information Sources in Neuroimaging: The NeuroBase Project

C. Barillot[1], H. Benali[2], M. Dojat[3], A. Gaignard[1], B. Gibaud[1], S. Kinkingnéhun[2], J-P. Matsumoto[4], M. Pélégrini-Issac[2], E. Simon[5], L.Temal[1]

Abstract: The NeuroBase project has for objective the study of the conditions for federating, through Internet, information sources in neuroimaging, where sources are distributed in different experimental sites, hospitals or research centers in cognitive neurosciences, and contain heterogeneous data and image processing programs. More precisely, this project consists in the creation of a shared ontology, suitable for supporting various neuroimaging applications, and a computer architecture for accessing and sharing relevant distributed information. We briefly describe the semantic model and report in more details the architecture we chose, based on a mediator/wrapper approach. To give a flavor of the future deployment of our architecture, we describe a demonstrator that implements the comparison of distributed image processing tools on distributed neuroimaging data.

Keywords: Medical Image Data bases, Mediation Systems, Mediator/Wrappers, Neuroimaging, Semantic Web, Medical Ontology

[1] Visages U746, INSERM-INRIA-CNRS-UnivRennes1, IRISA, Rennes, France
[2] IFR 49, CHR La Pitié Salpetrière/CEA-SHFJ, Paris, Orsay, France
[3] Unité INSERM U594, Grenoble, France
[4] Médience SA, INRIA, Rocquencourt, France

1 Introduction

One objective of neuroscientists is the construction of functional cerebral maps under normal and pathological conditions. Researches are currently performed to find correlations between anatomical structures, essentially sulci and gyri, where neuronal activation takes place, and cerebral functions, as assessed by recordings obtained by the means of various neuroimaging modalities, such as PET (Positron Emission Tomography), fMRI (functional Magnetic Resonance Imaging), EEG (Electro-EncephaloGraphy) and MEG (Magneto-EncephaloGraphy). Formation of such maps requires the development of sophisticated image processing techniques, such as segmentation and modeling of anatomical structures, registration and multimodality fusion, and specific methods for longitudinal data analysis.

Two of the major concerns of researchers and clinicians involved in neuroimaging experiments are on one hand, to manage internally the huge quantity of produced data ($\approx$1 Gb per subject) and, on the other hand, to be able to confront their experiences and the programs they develop, with those existing in other centers or moreover with those described in publications. Furthermore, and this is more particularly true for medium size centers (with limited staff capabilities), or even small ones (it is mostly the case in clinical centers), the researchers or the clinicians have great difficulties to set up large-scale experiments, mainly because of lack of man power and capacities of recruiting subjects. Besides, the statistical validity of the results is sometimes insufficient (the rate of "false negative" is probably not negligible). For all these reasons, we can think that the pooling of the experimental results, through a network between collaborative centers, will widen the scientific achievement of the conducted experimental studies. Through distributed neuroimaging data bases, the search for similar results, the search for images containing singularities or transverse searches via data mining techniques could highlight possible regularities. Moreover, this will also increase the possible panel of people involved in neuroimaging studies, while protecting the excellence of the supplied work.

In this context, NeuroBase is a cooperative project which aims to establish the conditions allowing, through Internet, the federation of distributed information sources in neuroimaging, these sources being located in various centers of experimentation, clinical departments in neurology, or research centers in cognitive neurosciences.

It requires that the users can diffuse, exchange or reach neuroimaging information with appropriate access means, in order to be able to retrieve information almost as easily as if it were stored locally.

1.1 Background

Because of the explosion of data generated by the neurosciences community, early in 90's has appeared the imperative necessity for innovative techniques for data and knowledge sharing and reuse [1,2]. This led to the starting of the North American ambitious "Human Brain Mapping" project. An objective recently added to this project is the development of data analysis and data processing software to operate on various data repository systems for data mining and knowledge discovery purposes. In parallel the development of web applications has stimulated the interest of researchers for distributed data bases and information sharing.

Four research topics are particularly relevant for our project:

1. Digital and probabilistic atlases of brain

   To gather and share neuroimaging information in a common referential space, various research efforts are performed for the construction of digital atlases: based on the labeling of post-mortem brains to quantify individual anatomical variability of cortical regions [3], for the anatomy and brain functions of rats [4] or of the primate visual system [5], or to associate symbolic data and graphical data about the nervous system [6]. Some atlases are developed to support interpretation of functional data [7], image processing instantiation in a specific context [8] or training [9].

For probabilistic atlases, some 300 MRI brain scans plus post morten data of 30 subjects were mixed in a common referential by the International Consorting for Brain Mapping [10]. Several image processing tools were added to allow segmentation and mapping of brain images to this brain reference.

2. Conception of image processing tools
The BRAID[5] project at Hopkins University is relevant here. It explores the anatomy-function relationship based on activation-response experiments and deficit-lesion analysis. The proposed system integrates mechanisms for complex queries, combining selection with multiple criteria, images quantification, and statistical tests to calculate correlations between deficits and lesions. Group studies rely on matching all brains to a target (reference) by the means of linear or non-linear (deformable elastic model) matching methods, each with its own pro and con. Several participants of this project have a well-known experience on the conception of such robust image processing tools.
3. Multi-center data bases
Several laboratories belonging to the Illinois University participate to the constitution of a commonly shared database devoted to neuronal patterns recordings. This work, oriented to animal recordings, is close to our project. The database is used for instance, to find temporal series specific to neurons populations under various stimuli conditions. A common data model has been developed to organize the experimental data. An atlas is available to enter, search and analyze heterogeneous data in a common referential. Ontology sharing and data schemata updating facilities have been also explored in the context of cooperative federated databases in [11].
4. Infrastructures for sharing data and processing tools
Several projects such as IXI [12] or Mammogrid [13] explore how grid technology can be applied to the field of medical image analysis by using large collections of computer resources to facilitate and scale processing across sites. The architectures proposed allow that image processing algorithms to be exposed as Grid services with the ability to compose these services as complex workflows executed across distributed resources. The notion of pipelines for the sequencing of image processing algorithms is also present in the LONI [14] or BrainVisa [15] frameworks.

2 The NeuroBase Approach

Instead of gathering all data in a central database [16], NeuroBase promotes a federated system for the management of distributed and heterogeneous information sources. The goal of the system is to allow the sharing of two types of information: on the one hand neuroimaging data, typically results from neuroimaging experiments, on the other hand image processing programs, typically image processing programs or statistical tools, being applied to the data available in the distributed system. Data can then be stored in relational databases or just in local files (wrappers will find their own way to the information). Image processing programs are modeled by the use of data flows. A data flow specifies the inputs and the parameters required for the completion of a given processing method, and the outputs of this procedure. Then, one of the most important aspects in this project is to define the main concepts shared by the different information centers in order to define a common semantic model every site can subscribe to (see Figure 1). From this base line, each site participating to the federated system can map its own concepts, data, image processing programs and ontologies, to this semantic referential [11]. For this purpose, we rely on a mediator/wrapper approach [17], where both the integration of (i) anatomical and functional images and related data (e.g. experimental protocols or subjects, pathology) and (ii) image processing programs, which can be applied to the images, (e.g. segmentation, registration, statistical analysis, …) can be expressed.

[5] BRAin Image Database, http://braid.rad.jhu.edu

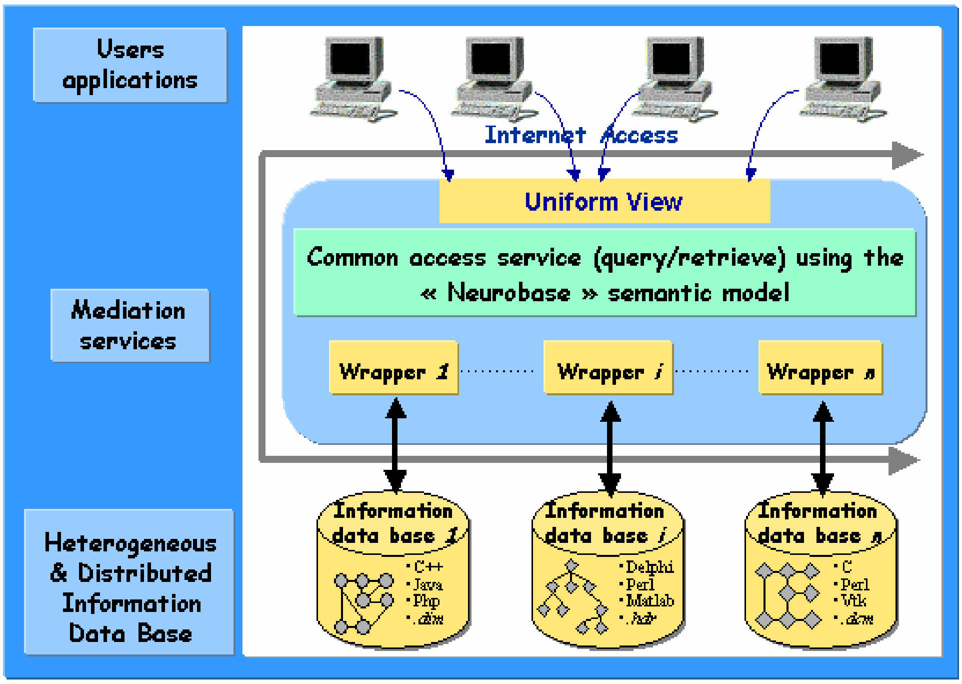

Figure 1: The NeuroBase system architecture for managing distributed information sources in neuroimaging. Mediation services are used to map and retrieve local information stored in heterogeneous and distributed databases following user queries expressed using concepts from the shared ontological model.

2.1 The Mediator/Wrapper approach

Mediators, systems for the mediation of information, were introduced to allow the virtual integration of heterogeneous distributed information sources in cooperative federated database systems. Mediators differ from standard database management systems in several aspects. Firstly, they do not supply mechanisms for simultaneous information sources updating. They only support queries to information sources in order to preserve their autonomy and the fact that they are locally managed. Secondly, to reinforce interoperability, to be highly adaptable to data structures encountered in databases, mediators support various data models from standard structured data, such as relational, object or multi-dimensional models, to semi-structured models, such as XML. The architecture of mediators is also different, based on a "mediator/wrapper" concept [17], in which a mediator offers a central view about all information sources and the associated wrappers dedicated to each source, hide their heterogeneity. A mediator redefines, using the corresponding wrappers, the user query into source dependent queries, then recomposes the various responses and formats the final response to the user. The query redefinition in sub-queries is optimized by the means of a cost based model to obtain the most efficient execution plan. This architecture clearly specifies the respective role of the mediator, which processes the user queries, and the wrappers, which translate the sub-queries into the relevant format for the associated information source. The pragmatic interest of such architecture is to lower the amount of work linked to the introduction of a new information source to the creation of the corresponding wrapper. Several mediators have been already developed (for instance DISCO[18], et Mocha [19]). Since 1998, one of the project's participant develops a new generation of mediators, called Le Select[6], which allows the sharing of distributed, heterogeneous

[6] http://www.medience.fr/index_en.html

and autonomous data and programs via a high level query language. Le Select is the corner stone of the NeuroBase approach.

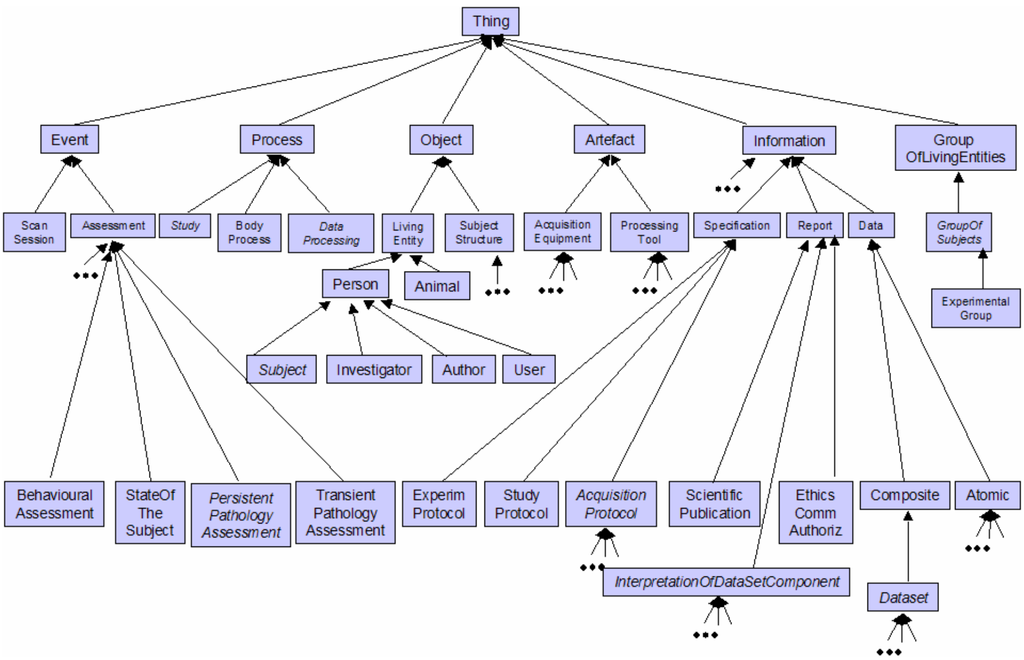

Figure 2. Excerpt of the NeuroBase ontology. Some concepts that appear in the text are shown in italic.

2.2 The Semantic Model

This semantic model or ontology has to be defined by a collaborative community, this requires quite a lot of work since there is nowadays not a fully defined common ontology from which we can derive our semantic referential. We have to build it in a domain which is complex and not well defined. Some existing works can provide valuable input such as the fMRI data center ontology[7] and medical thesauri such as the “Neuronames” terminology [20]. Our efforts related to the design of this ontology are detailed in a companion paper. Briefly, the concepts the ontology is made of, which represent the relevant entities and their associated properties, supply the search criteria susceptible to support user queries, such as a *Subject* or a *GroupOfSubjects* with or without a specific *PersistentPathologyAssessment* involved in a *Study*. Corresponding *Dataset* of Anatomical and Functional images with their *AcquisitionProtocol*, *DataProcessing* methods and *InterpretationOfDatasetComponent* (e.g. labels corresponding to anatomical entities, mesh, probabilistic information …) are described. Concepts have been introduced to cover at least specific applications addressed by the NeuroBase contributors (epilepsy, visual cortex exploration and Alzheimer disease).

From this semantic referential, contributors will have to subscribe to it and so forth to declare their own views, by means of site specific wrappers. They will determine what will be shared and what will not.

[7] http://www.fmridc.org/f/fmridc/aboutus/index.html

3 The NeuroBase Demonstrator

In order to evaluate our architecture, we have recently built a demonstrator based on some existing modules like Le Select, BrainVISA/Anatomist[8], MRIcro[9], FSL[10] and Vistal[11]. This can be extended to modules largely used in neurosciences communities such as SPM[12] software.

3.1 The test-bed application

The purpose of the test-bed application is to demonstrate that the NeuroBase architecture can support, via internet, the test and comparison of image processing modules in order for instance to select the most robust. These modules are distributed in several centers and applied on distributed data. Presently, two test centers are involved. A center, C1, located in Grenoble (FR) has developed an image processing chain for the delineation of visual cortical areas including cortex segmentation and unfolding. Image data are acquired on a Bruker scanner at 3T in the context of cognitive experiments for visual cortex exploration. They are stored using the Analyze format. A second center, C2, located in Rennes (FR) has developed image processing for restoration (denoising and debiasing) and segmentation. Data are mainly acquired in the context of eEpilepsy on a GE scanner at 1.5T and stored using the GIS format.
The schema in Figure 3 illustrates the application. First, C2 queries for an anatomical image available at C1 that is locally restored (anisotropic filtering) –i.e. at C2 - after the required format transformation. Then, C2 launches a specific tool for brain extraction. The Bet/FSL algorithm is executed at C1 on the input (a restored image) and provides the corresponding outputs: a brain image and a brain mask (binary image). After format conversion C2 fires locally the tissues segmentation. C2 launches a similar image processing available at C1 (MA_segmentation). Execution is then performed at C1. The two segmented images are then compared using the tool required (difference) at C2. Results are displayed at C2 or at C1 with the local 3D viewer. The same data flow can be executed on data from C1 (as in the example) or from C2. The final user does not need to know where the data are stored and where the methods are executed.

[8] http://brainvisa.info
[9] http://www.psychology.nottingham.ac.uk/staff/cr1/mricro.html
[10] http:/www./fmrib.ox.ac.uk.fsl/
[11] http://www.irisa.fr/visages/software-fra.html
[12] http://www.fil.ion.ucl.ac.uk/spm/

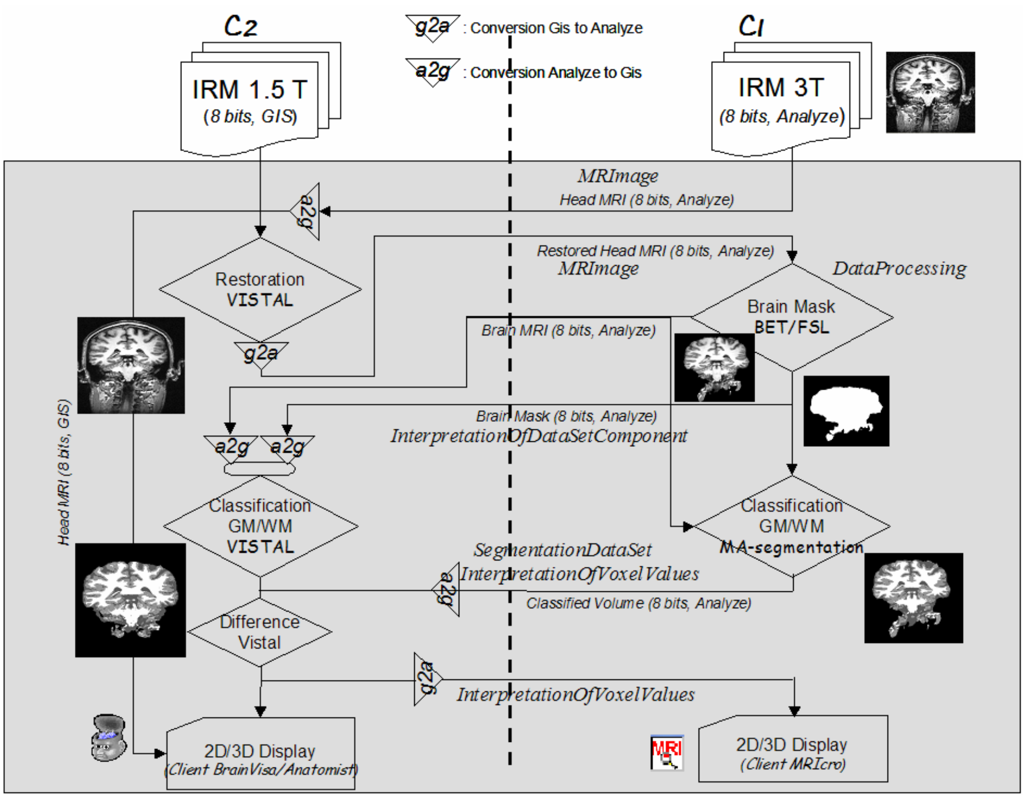

Figure 3: The test-bed application: two research centers C1 and C2, physically separated share, via internet, data and processing tools in order to compare two segmentation methods (Vistal in C2 and MA_segmentation in C1) on an anatomical image previously restored at C2. Segmentation processes are executed separately at each center. No synchronization process is implemented and difference is calculated when data are available. Concepts present in our ontology which correspond to inputs and outputs of our image processing tools are shown in italic.

3.2 The Architecture

The overall architecture is shown in Figure 4. The Le Select middleware is installed at each center. It is a generic server which includes data and image processing wrappers. Wrappers are site specific. Shared image processing tools are executed based on each local software library environment.

Local 2D/3D viewer (here Anatomist[13] and MriCro[14]) can be used. All distributed queries are performed via a common application developed in Tomcat servlet server environment accessible through a standard web browser.

[13] http://brainvisa.info/
[14] http://www.psychology.nottingham.ac.uk/staff/cr1/mricro.html

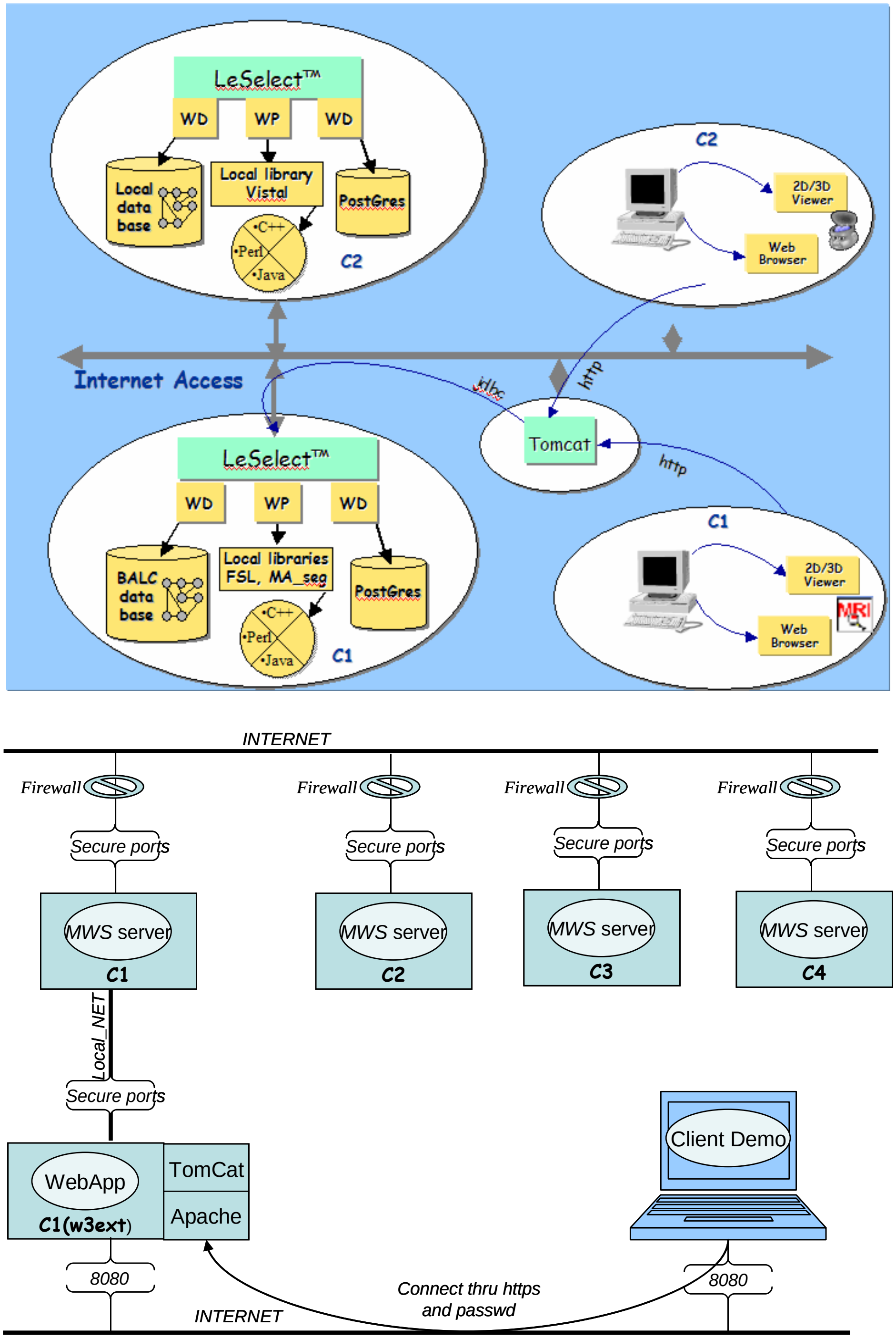

Figure 4: (*Top*) the NeuroBase demonstrator architecture for two distant centers C1 and C2. WD: data wrapper, WP: image processing program wrapper; (*Bottom*) the current network implementation of the system between 4 different partners, this underlines the generality of the proposed approach.

3.3 The Working Principles

Shared anatomical images are stored in local data repository (for C1 in a local file hierarchy and in a PostgreSQL database for C2). To make various queries using services available on our distributed system, wrappers were designed to map the local data organization in C1 and C2 with the semantic referential.

Figure 5 highlights the main mappings between local files hierarchy in C1 and some concepts.

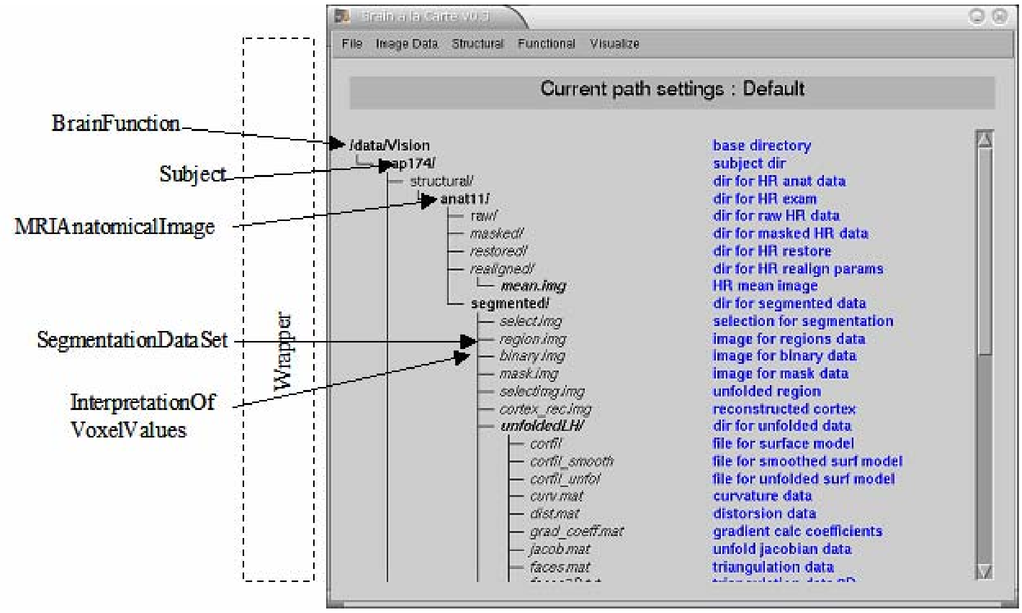

Figure 5: Mapping the BALC concepts hierarchy, ie the local database in center C1, to the semantic referential.

Similarly wrappers were introduced to execute programs on data published. As relational format is used by Le Select, program wrappers use relational data as input and relational data as output. In the following example, the skull stripping program is executed on images referred as Dataset in the ontology. This command is executed by Tomcat.

```
job execute //$host/fsl/Bet ← execution of bet program. host is set to C1 hostname
input a is
select Vol_Bin1 as img, Vol_Bin2 as hdr " + ← SQL query to retrieve input files /ontology/AllDataset
" +                          in AllDataset table of ontology
where ID = '$DatasetID'              ← identifier for all Dataset entities
```

4 Discussion

Our preliminary demonstrator shows that the principles and the technology we propose can be used in the context of neuroimaging data. Clearly, it should be extended. Presently, only a small part of the ontology is mapped to local data bases. The call of processing tools is hard-coded: selection of inputs and tuning parameters are still limited. The application developed in Tomcat environment should be extended to allow the selection, through a standard web browser, of the processing tools available. Finally, output of images processing tools available at each site are not reintroduced in the corresponding file hierarchy or PostgresSQL data base. Such extensions are under development. Neuroimaging is a relatively new scientific discipline in vivid evolution. Many concepts currently used were not present a few years ago. In this moving area, where no consensus is reached for several concepts, the definition of a centralized database for sharing data and processing methods seems rather complex and supposes, if succeeded, strong man power facilities for its maintenance. Moreover, there is a legitimate desire of autonomy that contradicts the centralized approach. Actually information sources exist in different centers but have generally been set up for purely local needs and are accessible to only a very small user community. In this context, the NeuroBase architecture we propose, based on a mediator/wrapper approach, seems attractive. Our architecture can be used to manage the evolution or even the upcoming of new information sources by just updating wrappers or creating new ones (this somewhat corresponds to the changing or the adding of new views to the semantic referential). In our approach, the semantic referential is central. It should be flexible enough to accept the introduction of new concepts, while keeping consistent. The AI community, from knowledge engineering to semantic grid has developed a strong expertise in this field via the construction of controlled vocabularies and thesauri that will provide valuable hints. The extensive use and the evolution of our demonstrator will allow us to confront it to different real situations.

5 References

[1] Mazziotta J.C., Toga A.W., Evans A.C., Fox P., and Lancaster J.L.: A Probabilistic Atlas of the Human Brain: Theory and Rationale for its development. Neuroimage 2 (1985) 89-10.

[2] Roland P.E., Zilles K.: Brain atlases - a new research tool. Trends in neurosciences 17(1994) 458-467.

[3] Graf von Keyserlingk, D., Niemann K., and Wasel, J.: A quantitative approach to spatial variation of human cerebral sulci. Acta Anatomica 131 (1988) 127-131.

[4] Toga, W.: A Three-Dimensional atlas of structure/function relationships, Journal of chemical neuroanatomy. 4 (1991) 313-318.

[5] Van Essen, D.C., Drury, H.A., Joshi, S. and Miller, M.I. Functional and structural mapping of human cerebral cortex: solutions are in the surfaces. Proc Natl Acad Sci U S A, 95 (1998) 788-795.

[6] Bloom, F.E.: The multidimensional database and neuroinformatics requirements for molecular and cellular neuroscience, Neuroimage 4 (1996) S12.

[7] Seitz, R.J., Bohm, C., Greitz, T., Roland, P.E., and Erikson, L.: Accuracy and precision of the computerized brain atlas programme for the localization and quantification in positron emmision tomography. J. Cerebr Blood Flow Metab. 10 (1990) 443-457.

[8] Lehmann, E.D., Hawkes, D.J., Hill, D.L., Bird, C.F., Robinson, G.P., Colchester, A.C. and Maisey M.N.: Computer-aided interpretation of SPECT images of the brain using an MRI-derived 3D neuro-anatomical atlas. Medical Informatics. 16 (1991) 151-66.

[9] Höhne, K.H., Bomans, M., Riemer, M., R. Schubert, R., Tiede, U. and Lierse, W.: A volume based anatomical atlas. IEEE Comp. Graphics and Application. 12 (1992)72-78.

[10] Tiede, U., Schiemann, T. and Höhne, K.H.: Visualizing the Visible Human. IEEE computer Graphics and Applications. 16 (1996) 7-9.

[11] Kahng J. and McLeod D. Dynamic Classificational Ontologies: Mediation of information sharing in Cooperative Federated Database Systems. In: Papazoglou M.P., Schlageter G. (eds): Cooperative Information Systems: Trends and Directions. Academic Press, London, U.K., (1998) 179-203.

[12] Rowland, A., Burns, M., Hartkens, T., Hajnal, J., et al. Information eXtraction from Images (IXI): Image Processing Workflows Using A Grid Enabled Image Database. In: Dojat M., Gibaud, B. (eds): Proceedings of DiDaMIC'04 Workshop. MICCAI conference (St Malo, 26-29 Sept. 04), Rennes (2004) 55-64.

[13] Brady, M. Grid-based Federated Databases of Mammograms: Mammogrid and eDiamond experiences. In: Dojat M., Gibaud, B. (eds): Proceedings of DiDaMIC'04 Workshop. MICCAI conference (St Malo, 26-29 Sept. 04), Rennes (2004) 84.

[14] Rex, D. E., Ma, J. Q. and Toga, A. W.: The LONI Pipeline Processing Environment. NeuroImage 19 (2003) 1033-48.

[15] Cointepas, Y., Mangin, J., Garnero, L., Poline, J., et al. BrainVISA: Software platform for visualization and analysis of multi-modality brain data. In: Proceedings of 7th Human Brain Mapping Conf, Brighton (UK) (2001) S98.
[16] Koslow, S.H.: Should the neuroscience community make a paradigm shift to sharing primary data? Nature Neuroscience. 3 (2000) 863-865.
[17] Wiederhold G. and Genesereth M.: The Conceptual Basis for Mediation Services. IEEE Expert 12 (1997) 38-47.
[18] Tomasic A. et al.. The distributed Information Search Component (Disco) and the World Wide Web. In Proc. of the ACM SIGMOD Int. Conf., Tucson, Arizona, May (1997) 546-548.
[19] Rodriguez-Martinez, M., Roussopoulos, N.: MOCHA: a Self-extensible Middleware Substrate for Distributed Data Sources, In: Proc. of the ACM SIGMOD Int Conf on Management of Data, Houston, May 2000.
[20] Bowden, D. M. and Martin, R. F.: NeuroNames Brain Hierarchy. NeuroImage 2 (1995) 63-83.